УДК 519.765:519.767:004.89

Б.М. Павлишенко
Львівський національний університет імені Івана Франка, Україна
Україна, 79005, Львів, вул. Драгоманова, 50

Використання лексемних полів у інтелектуальному аналізі текстових масивів

B.M. Pavlyshenko
Ivan Franko Lviv National University, Ukraine
Ukraine, 79005 Lviv, Drahomanov Str. 50

The Use of Lexemes Fields in Data Mining of Texts Arrays

Б.М. Павлышенко
Львовский национальный университет имени Ивана Франко, Украина
Украина, 79005, Львов, ул. Драгоманова, 50

Использование лексемных полей в интеллектуальном анализе текстовых массивов

У роботі запропонована модель семантичних та тематичних лексемних полів для інтелектуального аналізу текстових документів. Розглянуто векторну модель текстових документів у семантичному просторі, базис якого утворено частотно-дистрибутивними характеристиками семантичних та тематичних полів. Експериментальний аналіз тестової вибірки показав високу ефективність використання лексемних полів у класифікаційному аналізі авторства текстів.
Ключові слова: інтелектуальний аналіз даних, семантичні та тематичні поля, векторна модель текстових документів, класифікація текстів.

The model of semantic and thematic lexemes fields for data mining of text documents has been proposed. The vector model of text documents in the semantic space was considered. The basis of this space is formed by frequency-distributional characteristics of semantic and thematic fields. The experimental analysis of texts samples showed high efficiency of lexemes fields usage in the classification analysis of texts authorship.
Key words: data mining, Bayesian classification, semantic and thematic fields, vector space model of text documents, texts classification.

В работе предложена модель семантических и тематических лексемных полей для интеллектуального анализа текстовых документов. Рассмотрена векторная модель текстовых документов в семантическом пространстве, базис которого образован частотно-дистрибутивными характеристиками семантических и тематических полей. Экспериментальный анализ тестовой выборки показал высокую эффективность использования лексемных полей в классификационном анализе авторства текстов.
Ключевые слова: интеллектуальний анализ данных, семантические и тематические поля, векторная модель текстових документов, классификация текстов.

Вступ

Інтелектуальний аналіз текстових масивів є одним із перспективних напрямків сучасних інформаційних технологій. Складовими такого аналізу є алгоритми класифікації та кластеризації текстових документів. У цих алгоритмах використовують векторну модель текстових документів, яка базується на представленні документів як векторів у деякому фазовому просторі. Базис такого простору часто утворюють за допомогою частотно-дистрибутивних характеристик лексем текстового словника. Одна із основних проблем такого підходу зумовлена великою розмірністю аналізованого век-

торного простору. Також такий простір не дає можливості виділити задані семантичні складові в інтелектуальному аналізі текстів. У задачах аналізу текстового змісту актуальними є теорії лексичної семантики, зокрема, вчення про семантичні поля. Семантичні поля розглядають як групи лексем, об'єднаних спільним поняттям. Такі групи лексем утворюють нові характеристики текстових даних, використання яких може бути ефективним у задачах кластеризації та класифікації текстових документів. Семантичні поля глибоко вивчені у лінгвістичних працях, однак існує необхідність розробки формалізованих математичних моделей для їхнього впровадження в алгоритми інтелектуального аналізу текстових масивів.

Аналіз останніх досліджень та публікацій

У роботах [1], [2] описана векторна модель текстових документів. У [2-4] розглянуто методи класифікаційного аналізу текстових документів. У роботах [5-8] наведені результати аналізу текстових масивів на основі концепції семантичних полів. Семантичні поля розглянуті як групи лексем, об'єднаних спільним поняттям. У [5], [6] запропонована модель кластеризації текстових документів у семантичному просторі, яка дає можливість отримувати новий структурний поділ документів за семантичними ознаками у просторі суттєво меншої розмірності, ніж у просторі, утвореному частотними характеристиками лексемного складу текстової вибірки. У роботі [8] показано, що сингулярний розклад матриці семантичних ознак типу «частоти_семантичних_полів – документи» дає можливість аналізувати текстові документи у новому просторі семантичних концептів. Розглянемо лексикографічні концепції лексемних полів, які використовують у лінгвістиці. Семантичні групування слів відображають системність лексики. В основі визначення семантичних полів лежить лексико-семантична парадигма, під якою розуміють множину лексем, які об'єднані сукупністю семантичних ознак. Відмінність лексем у межах однієї парадигми визначається уточнюючими диференціюючими ознаками. Парадигми можуть бути одно- та багаторанговими. Ранги парадигми визначають структуру ієрархії лексемного об'єднання. Ядро семантичного поля утворюють лексеми, домінуюче значення яких визначають основними ознаками семантичного поля. Периферію семантичного поля утворюють лексеми, які містять основні поняття семантичного поля опосередковано, через ряд диференційних ознак, що мають відношення до основного поняття, яке утворює семантичне поле [9]. Одні і ті ж множини лексем називають як лексико-семантичні групи, семантичні поля, синонімічні ряди [10]. Уточнюючі та диференціюючі семантичні зв'язки в рамках одного семантичного поля визначають ієрархічну структуру поля [11]. Один із засновників вчення про семантичні поля – німецький вчений Трір, розділяв ієрархічну структуру лексем на словесні та понятійні поля. Він також вважав, що семантичні поля є неперервними, тобто лексеми семантичного поля охоплюють його понятійну область без пробілів так само, як склад словника охоплює весь спектр понять мови [12]. У лінгвістиці вводять поняття семантичного простору, який інтегрує та об'єднує семантичні поля [13]. На вершині семантичної організації знаходиться поняття семантичного простору, далі – поняття семантичного поля, лексико-семантичної групи, а на нижньому рівні знаходиться поняття слова. У роботі [14] введено поняття семантичних станів мовних одиниць, які розглянуті як формальні репрезентативні стани. У роботі [15] проаналізовані семантичні сітки, семантична структура та ієрархія лексичних одиниць. У роботі [16] запропонована концепція семантичних доменів, яка доповнює теорію семантичних полів. Визначення семантичних доменів є найбільш близьким до методів комп'ютерного аналізу текстів природної мови і базується на відповідних текстових колекціях, які належать до аналізо-

ваного домена і характеризують семантичні поняття, які виокремлюють аналізований домен. Лексемний склад семантичних полів визначають різними способами [17]. Один із способів полягає у виділенні загального поняття, на основі якого формують лексико-семантичне поле. Інший спосіб полягає у виділенні слова чи групи слів, до яких підбирають синонімічні ряди. Також виділяють семантичні поля на основі експертного аналізу спільних появ лексем у заданих контекстах. Прикладом комп'ютерної лексикографічної системи, в якій відображена семантична мережа зв'язків між лексемами, є система WordNet [18], яка розроблена у Прінстонському університеті. Ця система побудована на основі експертного лексикографічного аналізу семантичних структурних зв'язків, які відображають денотативні та конотативні характеристики лексемного складу словника. Глибина зв'язків у такій системі визначається експертною оцінкою лексемних комбінацій у текстових масивах і обмежується науковим досвідом експертів та об'ємом проаналізованого матеріалу. Семантичні поля у мережі WordNet представлені лексикографічними файлами. Іменники, дієслова, прикметники та прислівники організовані у синсети – множини синонімів. Іменники та дієслова згруповані відповідно до семантичних полів. У літературі розглядають такі лексемні класи, як семантичні поля, понятійні поля, тематичні групи лексем, семантичні групи, синонімічні ряди, семантичні домени та інші.

Підсумовуючи літературні дані досліджень семантичної класифікації лексемного складу словника можна побачити, що більшість визначень семантичної класифікації класів лексем є спорідненими, близькими до класичного визначення семантичного поля, і базуються на моделі «мішка слів». Відмінності між цими визначеннями зумовлені різним рівнем диференціації семантичних понять, на основі яких утворюють лексемні об'єднання. У цій моделі розглядають сукупність слів текстових документів без розгляду їх контекстуальної послідовності. На основі проаналізованого матеріалу можна зробити висновок про необхідність розробки комплексної структурної багаторівневої класифікаційної моделі лексемного складу текстових масивів, яка б об'єднувала на основі спільного теоретичного базису такі дистрибутивні лексемні відображення характеристик текстових масивів, як семантика документа, тематика масиву документів, семантична характеристика інформаційного джерела документів, характеристика авторів текстового масиву. Визначення об'єднуючого поняття семантичного поля потребує модельної та алгоритмічної формалізації. В залежності від обраної моделі та алгоритму об'єднання лексем можна отримати різні лексемні угрупування. На основі квантитативних характеристик кожного із таких угрупувань можна утворити додатковий вимір у семантичному просторі представлення текстових документів. Введення цих додаткових вимірів може бути ефективним у задачах інтелектуального аналізу текстів, зокрема у класифікаційних задачах та задачах кластерного аналізу. Велика розмірність векторного простору є значною проблемою класифікаційних алгоритмів. Тому актульними є методи зменшення розмірності базису. Структурування словника, зокрема у вигляді семантичної мережі, може дати суттєве зменшення розмірності базису внаслідок використання квантитативних ознак лексемних полів.

Постановка задачі

Побудуємо теоретико-множинну модель лексемних полів, яка буде описувати як лексико-семантичні, так і тематичні поля у лексемній структурі словників. Розглянемо модель текстових документів у просторі лексемних полів. Проведемо класифікаційний аналіз тестової вибірки текстових документів у просторі семантичних та тематичних полів. Як класифікатор оберемо наївний баєсівський класифікатор.

Теоретико-множинна модель лексемних полів

Розглянемо модель семантичних класів лексем, які утворені як на основі експертного лексикографічного групування лексем, так і на основі тематичних характеристик категоризованих текстових документів. Спочатку розглянемо модель класичного лексемного поля, яку в подальшому узагальнимо на випадок тематичного поля масиву категоризованих документів. Розглянемо утворення поняття «семантичне поле» в процесі аналізу текстових даних. Нехай існує деякий словник лексем, які зустрічаються в аналізованих текстових масивах. Опишемо цей словник як впорядковану множину

$$W = \{ w_i \mid i = 1,2...,N_w \}, \tag{1}$$

де N_w – кількість лексем у словнику. Введемо множину семантичних полів

$$S = \{ S_k \mid k = 1,2...,N_s \}, \tag{2}$$

де N_s – кількість семантичних полів. Семантичні ознаки лексем будемо характеризувати відображенням

$$U_{WS} : W \to S,\ w_i \to s_k,\ i = 1,2...,N_w; k = 1,2...,N_s. \tag{3}$$

Тобто у відповідність кожній лексемі ставлять деякий елемент множини *S*. Множина значень *S* може мати різну природу, наприклад, це може бути множина назв деяких семантичних класів. Шкала семантичних ознак є номінальною, якщо лексеми набувають деяких назв із множини *S*. Номінальна шкала володіє класифікаційним потенціалом, коли за допомогою відображення (3) можна утворити групування елементів множини *W*, які мають спільні назви із множини *S*. У загальному, класифікацію лексем за семантичними полями будемо розглядати як відображення множини лексем на множину семантичних полів. Семантичну класифікацію розглянемо як деяку сукупність відображень лексем на множину дійсних чисел. Можливу квантифікацію лексемних відображень можна пов’язати із частотами лексем у текстових об’єктах. Розглянемо утворення семантичного поля на основі відношення еквівалентності. Нехай існує деяке бінарне відношення

$$S_k^b \subseteq W \times W. \tag{4}$$

Розглянемо деяку квантитативну ознаку лексеми $x_k^s(w_i)$, яка кількісно характеризує лексемні відношення заданого типу у множині аналізованих текстових об’єктів. Наприклад, це може бути частота появи лексеми w_i в заданому лексемному шаблоні. Пов’яжемо із ознакою $x_k^s(w_i)$ бінарне відношення

$$S_k^b = \{ (w_i, w_j) | x_k^s(w_i) = x_k^s(w_j) \}. \tag{5}$$

Можна показати, що відношення S_k^b є рефлексивним, тобто

$$(w_i, w_i) \in S_k^b,\ \forall w_i \in W, \tag{6}$$

симетричним, тобто

$$(w_i, w_j) \in S_k^b \Rightarrow (w_j, w_i) \in S_k^b,\ \forall w_i, w_j \in W, \tag{7}$$

і транзитивним, тобто

$$(w_i, w_j) \in S_k^b, (w_j, w_l) \in S_k^b \Rightarrow (w_j, w_l) \in S_k^b,\ \forall w_i, w_j, w_l \in W. \tag{8}$$

Рефлексивне, симетричне і транзитивне відношення називають еквівалентністю [19]. Еквівалентність S_k^b повністю характеризує, породжуючи його ознаку, $x_k^s(w_i)$, і дає можливість визначити множину лексем, які не розрізняють за цією ознакою:

$$S_k^c = \{ w_i \mid (w_i, w_j) \in S_k^b \}. \quad (9)$$

Якщо S_k^c є деяким семантичним відношенням, тоді неспівпадаючі множини S_k^c утворюють розбиття лексемного словника *W* на семантичні класи

$$S_{sc} = \{ S_k^c \mid k = 1,2..., N_s \}. \quad (10)$$

Такі семантичні класи, враховуючи теорію лексико-семантичних полів, можна розглядати як лексемні поля. Бінарне відношення S_k^b може також породжуватись деяким логічним висловлюванням $Q(w_i, w_j)$

$$S_k^b = \{ (w_i, w_j) | Q(w_i, w_j) = true \}, \quad (11)$$

де $Q(w_i, w_j)$ описує деяку умову, наприклад, одночасне використання в текстових шаблонах заданої структури. Умова породження бінарного відношення S_k^b може також описуватись деяким правилом підстановки в заданій схемі формальної граматики. Таке правило може бути сформовано деяким регулярним виразом. Розглянемо рангову ознаку $x_k^{rs}(w_i)$, яка утворює бінарне відношення

$$S_k^{rb} = \{ (w_i, w_j) | x_k^s(w_i) \le x_k^s(w_j) \}. \quad (12)$$

Можна показати, що таке бінарне відношення є рефлексивне, транзитивне та лінійне. Такі відношення називають лінійними квазіпорядками [19]. Квазіпорядок S_k^{rb} породжує рангову шкалу семантичного поля S_k^r. У випадку формування семантичного поля за допомогою рангових ознак можна визначити внутрішню структуру поля, для якої можна сформувати внутрішній частковий порядок, виділивши структурні групи всередині семантичного поля. Такими групами можуть бути, наприклад, частотне ядро семантичного поля, основна частотна область, периферійна частотна область. Для кожної із цих груп можна визначити умови для семантичної ознаки, за якою лексеми всередині цих груп не розрізняють. Відношення еквівалентності та квазіпорядку визначають номінальні та рангові семантичні шкали для лексемного складу словника текстових масивів на основі лексемних відношень елементів різних класів семантичного розбиття.

Введемо поняття тематичного поля за аналогією із семантичним полем. Вважаємо, що тематичне поле утворюють лексеми словника текстових масивів, які характеризують тематику деякої категорії текстових документів. Такі категорії можна визначати, наприклад, на основі дистрибутивних характеристик текстів, згрупованих за деякою визначеною тематикою, авторством текстів, джерелом походження тощо. Множину тематичних полів позначимо так

$$Them = \{ them_i | 1,2,...N_{them} \}, \quad (13)$$

де $N_{them} = |Them|$ – розмір множини тематичних полів, який визначений кількістю тематичних категорій. Введемо деякий коефіцієнт, який буде відображати, у скільки разів деяку лексему вживають частіше у деякій категорії у порівнянні із

загальною вибіркою усіх категорій. Визначимо цей коефіцієнт як відношення частоти лексеми у документах заданої категорії до частоти цієї ж лексеми у загальній текстові вибірці

$$Kthem_{ij}^{wg} = \frac{p_{ij}^{wg}}{p_i^{w}}. \qquad (14)$$

Назвемо $Kthem_{ij}^{wg}$ коефіцієнтом тематичної виразності. Визначимо тематичне поле $them_k$ деякої категорії текстових документів ctg_k, як підмножину словника лексем, для яких коефіцієнт тематичної виразності є більший за деяке, наперед визначене, значення:

$$W_k^{them} = \{w_i \mid Kthem_{ik}^{wg}(w_i) > Kthem_t\}, \qquad (15)$$

де $Kthem_t$ – деяке порогове значення коефіцієнта тематичної виразності.

На основі визначення множини тематичного поля можна сформувати лексемний склад для кожного тематичного поля, заданого певною категорією текстових документів. Введення простору семантичних та тематичних полів не тільки зменшує розмірність задачі аналізу текстів, а також вводить новий базис для текстових характеристик.

У семантичному базисі можуть спостерігатися якісно нові групування текстових документів.

Розгляд таких групувань може бути ефективним в алгоритмах комплексного аналізу текстів.

Векторна модель текстових документів

Розглянемо формування базису лексемних семантичних та тематичних полів для векторного простору текстових документів.

Сукупність текстових документів опишемо такою множиною

$$D = \{ d_j \mid j = 0,1,2..., N_d \}, \qquad (16)$$

де N_d – кількість документів. Під документом з $j = 0$, будемо вважати документ з нейтральним текстом, який відповідає лінгвостатистичній нормі. Документ d_j з множини текстових документів D можна представити як упорядковану множину слів T_j^d, порядок елементів якої відповідає порядку слів у цьому документі:

$$T_j^d = \{ t_{lj} \mid l = 1,2...N_j^t \}. \qquad (17)$$

Упорядкований за алфавітом словник текстового документа d_j розглянемо як мультимножину W_j^d над множиною словника W

$$W_j^d = \{ n_{ij}^{wd}(w_i) \mid w_i \in d_j, i = 1,2..., N_w \}, \qquad (18)$$

де n_{ij}^{wd} – кількість входжень лексеми w_i із словника W в множину лексем текстового документа d_j, яку можна визначити як

$$n_{ij}^{wd} = \sum_{l=1}^{N_j^t} f_{wd}(t_{lj}, w_i), \quad f_{wd}(t_{lj}, w_i) = \begin{cases} 1, t_{lj} = w_i \\ 0, w_{lj}^d \neq w_i \end{cases}. \qquad (19)$$

Відображення лексемного складу словника W на множину семантичних полів S (3) задамо таблицею, яка визначена експертним лексикографічним аналізом. Лексемний склад семантичного поля s_k визначимо як

$$W_k^s = \left\{ w_i \mid w_i \xrightarrow{U_{ws}} s_k, i = 1,2..., N_w \right\}. \quad (20)$$

Множину образів відображення U_{ws} (3) розглянемо як мультимножину над множиною семантичних полів S

$$S_f = \left\{ n_k^s(s_k) \mid k = 1,2..., N_s \right\}, \quad (21)$$

де n_k^s – кількість лексем словника W, які відносяться до семантичного поля s_k :

$$n_k^s = \sum_{i=1}^{N_w} f_s(w_i, s_k), \text{ де } f_s(w_i, s_k) = \begin{cases} 1, w_i \in W_k^s \\ 0, w_i \notin W_k^s \end{cases}. \quad (22)$$

Введемо мультимножину образів відображення U_{ws} семантичних полів для окремого документа d_j

$$S_j^d = \left\{ n_{kj}^{sd}(s_k) \mid k = 1,2..., N_s \right\}, \quad (23)$$

де n_{kj}^{sd} – кількість лексем семантичного поля s_k в лексемному складі документа d_j

$$n_{kj}^{sd} = \sum_{l=1}^{N_j^t} f_s(t_{lj}, s_k), \text{ де } f_s(t_{lj}, s_k) = \begin{cases} 1, t_{lj} \in W_k^s \\ 0, t_{lj} \notin W_k^s \end{cases}. \quad (24)$$

Введемо оператор відображення лексемного словника W на множину квантитативних ознак у масиві документів

$$U_{wd} : w_i \to p_{ij}^{wd}, \; i = 1,2,..., N_w, j = 1,2..., N_d. \quad (25)$$

У загальному випадку величина p_{ij}^{wd} може мати довільне походження квантитативної характеристики.

У подальшому будемо розглядати цю величину як текстову частоту лексеми w_i у текстовому документі d_j, яка визначена такою функціональною залежністю

$$p_{ij}^{wd} = \frac{n_{ij}^{wd}}{N_j^t}. \quad (26)$$

Аналогічно введемо оператор відображення семантичного складу S_j^d текстового документа d_j на множину квантитативних ознак:

$$U_{sd} : s_k \to p_{kj}^{sd}, \; k = 1,2,...N_s, \; j = 1,2...N_d. \quad (27)$$

Величина p_{kj}^{sd} визначає структурну частоту лексем семантичного поля s_k у текстовому документі d_j. Визначимо p_{kj}^{sd} за такою формулою

$$p_{kj}^{sd} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_s(w_i, s_k), \text{ де } f_s(w_i, s_k) = \begin{cases} 1, & w_i \in W_k^s \\ 0, & w_i \notin W_k^s \end{cases}. \quad (28)$$

Сукупність значень p_{ij}^{wd} утворює матрицю типу ознака-документ

$$M_{wd} = \left(p_{ij}^{wd}\right)_{i=1,\,j=1}^{N_w,N_d}. \qquad (29)$$

У матриці M_{wd} роль ознаки відіграє текстова частота лексеми. Введемо вектор

$$V_j^w = \left(p_{1j}^{wd}, p_{2j}^{wd}, ..., p_{N_w j}^{wd}\right). \qquad (30)$$

Такий вектор відображає документ d_j в N_w-мірному просторі текстових документів. Сукупність значень p_{kj}^{sd} утворюють іншу матрицю ознака-документ, у якій ознаками виступають частоти семантичних полів у документах:

$$M_{sd} = \left(p_{kj}^{sd}\right)_{k=1,\,j=1}^{N_s,N_d}. \qquad (31)$$

Вектор

$$V_j^s = \left(p_{1j}^{sd}, p_{2j}^{sd}, ..., p_{N_s j}^{sd}\right) \qquad (32)$$

відображає документ d_j в N_s-мірному просторі текстових документів.

Текстові документи можуть бути представлені за допомогою тематичних векторів V_j^{them}, які визначають за аналогією до семантичних векторів.

Розглянемо поняття тематичного поля як сукупності лексем, які в загальному випадку можуть належати різним частинам мови і повинні однозначно відображати понятійний спектр деякої категорії текстових документів.

Аналогічно до частот семантичних полів визначимо частоти тематичних полів кожного документа як суми частот лексем, які належать цьому полю:

$$p_{kj}^{(them)d} = \sum_{i=1}^{N_w} p_{ij}^{wd} f_{them}(w_i, them_k), \quad f_{them}(w_i, them_k) = \begin{cases} 1, & w_i \in W_k^{them} \\ 0, & w_i \notin W_k^{them} \end{cases}, \qquad (33)$$

де $p_{kj}^{(them)d}$ – частота тематичного поля $them_k$ у текстовому документі d_j, W_k^{them} – множина лексем тематичного поля $them_k$, визначена формулою (15). Розглянемо матрицю $M_{(them)d}$ типу тематичні поля-документи за аналогією до матриці семантичних полів M_{sd}

$$M_{(them)d} = \left(p_{kj}^{(them)d}\right)_{k=1,\,j=1}^{N_{them},N_d}, \qquad (34)$$

де $p_{kj}^{(them)d}$ – частоти тематичних полів, N_{them} – кількість тематичних полів, N_d – кількість текстових документів. Частоти тематичних полів утворюють координати текстових повідомлень у векторному семантичному просторі. Вектор

$$V_j^{them} = \left(p_{1j}^{(them)d}, p_{2j}^{(them)d}, ..., p_{N_{(them)}j}^{(them)d}\right) \qquad (35)$$

відображає документ d_j в N_w-мірному просторі, базис якого утворений тематичними полями. Використання векторного представлення дає можливість пошуку подібних документів та псевдодокументів у векторному просторі із базисом, утвореним частотними характеристиками семантичних та тематичних полів. Цей базис має суттєво меншу розмірність у порівнянні із базисом, утвореним частотними характеристиками лексем словника текстових масивів. Це дає можливість зменшити кількість необхідних обчислень в алгоритмах аналізу текстів.

Експериментальні дослідження

Для експериментального вивчення класифікації текстових документів у просторі семантичних полів ми вибрали текстову базу 503 художніх творів 17 авторів. Для формування семантичного простору вибрано лексеми, згруповані за семантичними полями іменників та дієслів семантичної мережі WordNet [18]. Семантичні поля у мережі WordNet (http://wordnet.princeton.edu) представлені лексикографічними файлами. У наших дослідженнях ми використали семантичні поля іменників та дієслів. Семантичні поля іменників складаються із 26 лексикографічних файлів, із яких ми вибрали 54 464 лексеми. Семантичні поля дієслів містять 15 лексикографічних файлів, у які ми відібрали 9097 лексем. Також розглянуто 17 тематичних полів за тематичними категоріями текстових документів, згрупованих за авторами. Коефіцієнт тематичності, за яким відібрані лексеми для тематичних полів, був більшим за мінімальне значення, що дорівнює 2. Тобто тематичні поля для категорії текстів деякого автора сформовані на основі лексем, які зустрічаються у цих текстах у два і більше разів частіше, ніж у сукупній вибірці текстів усіх авторів. Навчальна вибірка містила 350 документів, а тестова – 153. Для класифікації текстових документів вибрано наївний баєсівський класифікатор. Класифікація текстових документів у просторі семантичних полів за допомогою баєсівського класифікатора описана в [8]. Для характеристики класифікаторів використовують поняття точності (precision) та повноти (recall) [3], [4]. Точність класифікатора $\Pr_j$ для категорії Ctg_j визначають як відношення кількості елементів, які правильно класифіковані як належні до категорії Ctg_j до загальної кількості елементів, які класифіковані як належні до категорії Ctg_j

$$\Pr_j = \frac{\left|\left\{d_i \mid Class(d_i) = Ctg_j \wedge d_i \in Ctg_j\right\}\right|}{\left|\left\{d_i \mid Class(d_i) = Ctg_j\right\}\right|}, \tag{36}$$

де $Class(d_i)$ – визначена класифікатором категорія. Повноту (recall) класифікатора Rc_j визначають як відношення успішно класифікованих документів у заданій категорії до загальної кількості документів у цій категорії.

$$Rc_j = \frac{\left|\left\{d_i \mid Class(d_i) = Ctg_j \wedge d_i \in Ctg_j\right\}\right|}{\left|\left\{d_i \mid d_i \in Ctg_j\right\}\right|} \tag{37}$$

Розглянемо основні отримані результати. Для класифікатора у просторі семантичних полів отримано такі значення точності та повноти класифікації: $Pr_{mean}^{tclass} = 0.7066$, $Rc_{mean}^{tclass} = 0.6952$. При тестовій класифікації документів за авторами у просторі тематичних полів отримано такі значення точності та повноти класифікації: $Pr_{mean}^{tclass} = 0.914$, $Rc_{mean}^{tclass} = 0.898$. Графік розподілу точності та повноти баєсівського класифікатора у просторі тематичних полів наведено на рис. 1.

Як випливає із отриманих результатів, представлення текстів у просторі семантичних та тематичних полів дає високі результати точності класифікаційного аналізу авторства текстів для розглянутої текстової вибірки художніх творів.

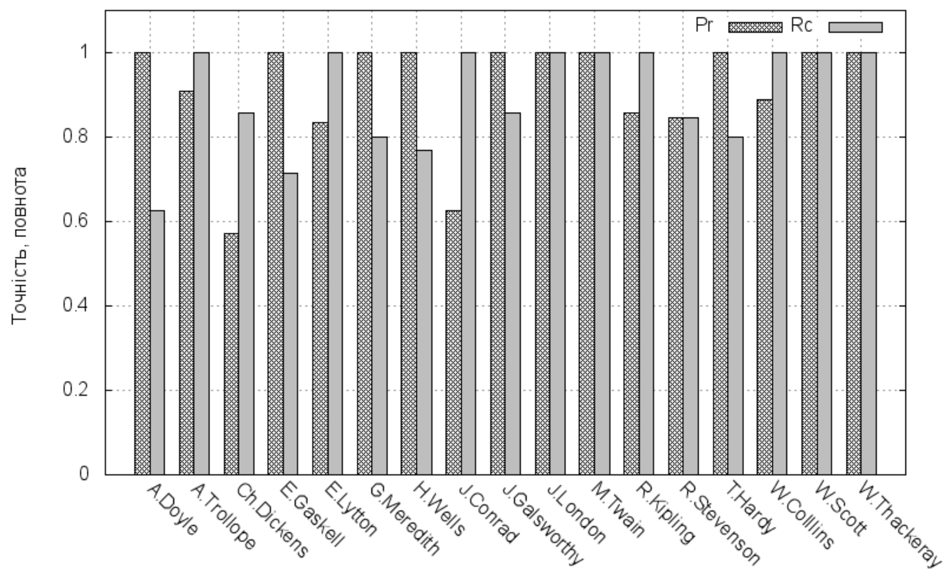


Рисунок 1 – Розподіл точності та повноти для баєсівського класифікатора у просторі тематичних полів

Висновки

У роботі розглянуті лінгвістичні концепції семантичних та тематичних лексикографічних полів із точки зору їх використання в алгоритмах інтелектуального аналізу текстових масивів. Під семантичними полями розглядають множини лексем, які об’єднані деякою парадигмою. Під парадигмою можна розуміти, наприклад, спектр семантичних або тематичних понять, які відображені у структурі лексикографічних значень лексем. На основі концепцій семантичних полів створена теоретико-множинна модель, яка об’єднує поняття семантичного та тематичного лексемного поля. Лексикографічні семантичні та тематичні поля можна розглядати як підкласи об’єднуючого класу лексемних полів. Лексемні поля розглянуті як розбиття лексемного словника на основі відношення еквівалентності. Лексикографічні поля утворені на основі експертного семантичного групування лексемного складу словника. Тематичні поля утворені на основі лексем, які характерні для тематично категоризованих текстових документів і визначаються на основі коефіцієнта тематичної виразності. Цей коефіцієнт показує, у скільки разів лексеми тематичного поля зустрічаються частіше у текстах заданої тематичної категорії у порівнянні із текстами лінгвостилістичної норми. Розглянуто векторну модель текстових документів у семантичному просторі, базис якого утворено частотно-дистрибутивними характеристиками семантичних та тематичних полів. Експериментальний класифікаційний аналіз тестової вибірки текстових документів у векторному просторі семантичних та тематичних полів показав високу ефективність використання лексемних полів у класифікаційному аналізі. Точність наївного баєсівського класифікатора у просторі тематичних полів для проаналізованої вибірки авторських текстів є вищою у порівнянні із такою ж точністю у просторі лексикографічних семантичних полів. Базис лексикографічних семантичних полів є незалежним від вибірки, а базис тематичних полів є індивідуальним для кожної текстової вибірки.

Література

1. Pantel P. From Frequency to Meaning: Vector Space Models of Semantics/ Pantel Patrick, Turney Peter D. // Journal of Artificial Intelligence Research. – 2010. – Vol. 37. – P. 141-188.
2. Брасегян А.А. Анализ данных и процессов : [учеб. пособие] / А.А. Брасегян, М.С. Куприянов, И.И. Холод [и т.д.]. – СПб. : БХВ-Петербург,2009. – 512 с. : ил.
3. Sebastiani F. Machine Learning in Automated Text Categorization / F. Sebastiani // ACM Computing Surveys. – 2002. – Vol. 34, № 1. – P. 1-47.
4. Manning C.D. Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval / Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. – Cambridge University Press, 2008. – 496 p.
5. Павлишенко Б.М. Ієрархічна кластеризація текстових документів у векторному просторі семантичних полів / Б.М. Павлишенко // Електроніка та інформаційні технології. – 2011. – Випуск 1. – С. 212-222.
6. Павлишенко Б. Семантична кластеризація текстових документів методом k-середніх / Б. Павлишенко // Комп'ютерні науки та інформаційні технології : збірник наукових праць. – Львів : Видавництво Львівської політехніки. – 2011. – № 710. – С. 215-218.
7. Павлишенко Б.М. Сингулярна декомпозиція матриці семантичних ознак в алгоритмі ієрархічної кластеризації текстових масивів / Б.М. Павлишенко // Математичні машини і системи. – 2012. – № 1. – С. 69-76.
8. Павлишенко Б.М. Ймовірнісна класифікація текстових документів в просторі семантичних полів / Б.М. Павлишенко // Електроніка та інформаційні технології. – 2012. – Випуск 2.– С. 164-172.
9. Вердиева З.Н. Семантические поля в современном английском языке / Вердиева З.Н. – М. : Высшая школа, 1986. – 120 с.
10. Полевые структуры в системе языка : [коллективная монография] / [под. ред. проф. З.Д. Попова]. – Воронеж : Изд-во Воронежского ун-та, 1989. – 197 с.
11. Лексико-семантические группы русских глаголов / [под. ред. Э.В. Кузнецовой]. – Иркутск : изд. Иркут. ун-та, 1989. – 180 с.
12. Уфимцева А.А. Опыт изучения лексики как системы (на материале английского языка) / Уфимцева А.А. – М. : Издательство Академии наук СССР, 1962. – 176 с.
13. Русанівський В.М. Інформаційно-лінгвістичні основи тлумачної лексикографії / В.М. Русанівський, В.А. Широков // Мовознавство. – К., 2002. – № 6. – С. 7-31.
14. Широков В.А. Семантичні стани мовних одиниць та їх застосування в когнітивній лексикографії / В.А. Широков // Мовознавство. – 2005. – № 3-4. – С. 47- 62.
15. Скороходько Е.Ф. Сіткове моделювання лексики: лінгвістична інтерпретація параметрів семантичної складності / Е.Ф. Скороходько // Мовознавство. – 1995. – № 6. – С. 19-28.
16. Gliozzo A. Semantic Domains in Computational Linguistics / Alfio Gliozzo, Carlo Strapparava. – Springer, 2009. – 132 p.
17. Гольдберг В.Б. Контрастивный анализ лексико-семантических груп (на материале английского, русского и немецкого языков) / В.Б. Гольдберг. – Тамбов : ТГПИ, 1988. – 56 с.
18. Fellbaum C. WordNet. An Electronic Lexical Database / Fellbaum C. – Cambridge, MA : MIT Press, 1998. – 432 p.
19. Миркин Б.Г. Анализ качественных признаков и структур / Миркин Б.Г. – М. : Статистика, 1980. – 319 с., ил.

Literatura

1. Pantel P. Journal of Artificial Intelligence Research. 2010. Vol.37. P.141-188.
2. Brasegyan A.A. Analiz dannyh i protsessov: ucheb. posobie. SPb.:BHV-Peterburg,2009. 512s.
3. Sebastiani F. ACM Computing Surveys. 2002. Vol. 34. № 1. P. 1-47.
4. Manning C. D. Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval. Cambridge University Press. 2008. 496p.
5. Pavlyshenko B. M. Elektronika ta informatsiyni tehnologii. 2011. Vypusk 1. S. 212-222.
6. Pavlyshenko B. Komp'yuterni nauky ta informatsiyni tehnologii : zbirnyk naukovyh prats'. L'viv : Vydavnystvo L'vivs'koi politehniky. 2011. № 710. S. 215-218.
7. Pavlyshenko B. M. Matematychni mashyny i systemy. 2012. №1. S. 69-76.
8. Pavlyshenko B.M. Elektronyka ta informatsiyni tehnologii. 2012. Vypusk 2 . S.164-172.
9. Verdieva Z.N. Semanticheskie polya v sovremennom angliyskom yazyke. M.: Vysshaya shkola. 1986. 120s.
10. Polevye struktury v sisteme yazyka./kollektivnaya monografiya pod.red. prof. Z.D.Popova. Voronezh.: Izd-vo Voronezhskogo un-ta.1989. 197s.

11. Kuznetsova E. V. Leksiko-semanticheskie gruppy russkih glagolov. Irkutsk: Izd-vo Irkut. Un-ta. 1989. 180s.
12. Ufimtseva A.A. Opyt izucheniya leksiki kak sistemy (na materiale angliyskogo yazyka). M.: Izdatel'stvo Akademii nauk SSSR. 1962. 176s.
13. Rusanivs/ky V.M. Informatsiyno-lingvistychni osnovy tlumachnoi leksykografii. Movoznavstvo. K. 2002. №6. S.7-31.
14. Shyrokov V.A. Semantychni stany movnyh odynyts' ta ih zastosuvannya v kognityvniy leksykografii. Movoznavstvo. 2005. №3-4. S.47- 62.
15. Skorohod'ko E.F. Sitkove modeluvannya leksyky: lingvistychna interpretatsiya parametriv semantichnoi skladnosti. Movoznavstvo. 1995. №6. S.19-28.
16. Gliozzo A. Semantic Domains in Computational Linguistics. Alfio Gliozzo, Carlo Strapparava. Springer. 2009. 132 p.
17. Gol'dberg V.B. Kontrastivnyj analiz leksiko-semanticheskih grup (na materiale angliyskogo, russkogo i nemetskogo yazykov). Tambov: TGPI. 1988. 56 s.
18. Fellbaum C. WordNet. An Electronic Lexical Database. Cambridge. MA: MIT Press. 1998. 432 p.
19. Mirkin B.G. Analiz kachestvennyh priznakov i struktur. M.: Statistika. 1980. 319 s.

RESUME

B.M. Pavlyshenko

The Use of Lexemes Fields in Data Mining of Texts Arrays

This paper describes the linguistic concepts of semantic and thematic lexicographical fields in terms of their use in the algorithms of text arrays data mining. Semantic fields are the set of lexemes which are united under some paradigm. The paradigm can be, for example a range of semantic or thematic concepts which are represented in the structure of lexemes lexicographical value. On the basis of the semantic fields concepts we created a set-theoretical model which combines the concepts of semantic and thematic lexeme fields. Lexicographic semantic and thematic fields may be considered as subclasses of a unifying class of lexeme fields. Lexeme fields are considered as a set partition of a lexeme dictionary based on the equivalence relation. Lexicographic fields are formed on the basis of expert semantic grouping the dictionary lexeme structure. Thematic fields are created from the lexemes typical for thematically categorized text documents and are determined due to the coefficient of thematic expressiveness. This coefficient shows how many times the lexemes of thematic fields are more frequent in the texts of given thematic category as compared to the texts of linguo-stylistical norm. We also studied a vector model of text documents in the semantic space, the basis of which is formed by frequency-distributional characteristics of semantic and thematic fields. Experimental classification analysis of the test sample of text documents in the vector space of semantic and thematic fields showed high effectiveness in using lexeme fields for classification analysis. The precision of naive Bayesian classifier in the space of thematic fields is higher for analyzed authors' texts in comparison with the same precision in the space of lexicographic semantic fields. The basis of lexicographic semantic fields is independent of the texts sample, the basis of thematic fields is specific to each texts sample.

Стаття надійшла до редакції 07.11.2012.